CKME136 – Capstone Project Literature Review

Comparative Analysis for Walmart Sales Forecast

Adrian Wong (501030070)

2020-06-09

Comparative Analysis for Walmart Sales Forecast

Introduction

Walmart Inc. is a multinational retail company that has more than ten-thousands stores worldwide. Selling all sorts of products including, grocery, household items, furniture, clothing, jewelry, electronics, and more. It is always hard to predict accurate sales for a huge company like Walmart, as many hidden factors could affect sales. Since the past centuries, various predictive methods and models have been developed, and Walmart Inc. must choose an accurate method for predicting future sales as it influences management decisions. With an accurate prediction of sales, management could better allocate resources for marketing and finding out which departments are doing poorly to maximize profit. Using Walmart Dataset from Kaggle as a case study, this project seeks to address the following question:

With the use of deep learning approaches in time series data, does the predict of sales become more accurate compared to traditional time series methods and regression methods?

By computing time series methods (Moving Average - MA, Auto-Regressive Integrated Moving Average - ARIMA, Seasonal ARIMA - SARIMA, Holt's Winter seasonal method, and Support Vector Regression - SVR), deep learning of neural network methods (Long Short Term Memory - LSTM and Gated Recurrent Units - GRU), also with regression methods (Univariate Linear Regression, Multiple Regression, Decision Tree and Random Forest). All of the mentioned above methods would be compared against Weighted Mean Absolute Error (WMAE) as requested in Kaggle. Tools such as RStudio, Python, Weka, and Tableau would be used throughout the project. The best model would be used to compute future sales in the test.csv as a conclusion. Limitations and recommendations will also be provided.

Literature Review

Businesses have been trying to utilizing math and different technique to predict sales, with accurate prediction, it is very helpful for them to prepare how much to order and for marketing. Throughout the centuries, many methods are being developed and used for predictions. It is always a hard choice to choose what methods as each technique has its strength and limitation (Mentzer & Moon, 2004) (Lasek et al., 2016).

Timeseries and Hybrid Methods with Neural Network:

When there is uncertainty in predicting future values that are given with time stamp, the time-series approach seems to work the best (Goyall et al., 2018). It is not surprising that a researcher found out that more than 50% of the research papers on predictive analytics for the supply chain used the time series forecasting approach. (Bonnes, 2014). By using traditional simple time series methods such as Autoregression (AR) and MA, a lot of the times the accuracy is not as desirable (Udom & Phumchusri, 2014). Researchers have stepped up and combining or modifying existed methods, such as developed SARIMA and SVR, which shows better results (Pai & Lin, 2005) (Pai et al., 2010). In the recent decade, researchers utilized unsupervised machine learning and deep learning in time series modelings such as LSTM and GRU. Even though the main weakness of LSTM and GRU are their explainability, --as there is not much justification provided for these models (Baccar et al., 2019). Most researchers suggested that if the methods are trained and configurated properly, it yields better results than traditional approaches (Längkvist et al., 2014) (Bandara et al., 2019).

Regressions and Hybrid Methods:

On the other hand, some researchers argue that sales prediction is rather a regression problem, this is because the time-series approach analyzes time series sequences and investigate what are the parameter statistics and other characteristics, and used that to predict future values using historical observed value. Whereas regression, it is more like testing theories, checking whether the current value in one or more independent time series influences the current value in another time series (Pavlyshenko, 2019). Similar to the time-series approach, the simpler regression models such as Ordinary Least Squares (OLS) is never the best methods for predicting sales. More sophisticated models like multiple linear regression, decision tree, random forest, and XGBoost outperform simple regression models (Wu et al.,2018) (Hülsmann et al., 2012) (Jain et al., 2015).

Comparison of both approaches:

Regardless of using the time-series approach or regression approach, one thing for sure is that there are never the “best” methods, as each dataset is different, and most of these methods and techniques have their strengths and weaknesses. Testing and comparing the results are needed to see which yields the best results. Therefore, in this research paper, the main focuses are to determine which methods work the best with the Walmart Sales forecast and would be ranked based on WMAE.

Dataset

In the Kaggle dataset, Walmart has provided historical weekly sales data extracted from the 45 Walmart store from 05/02/2010 to 26/10/2012, a total of 421,570 instances. Each of the stores contains around 90 departments. Additional to date and weekly sales amount, the dataset also included other variables such as store size, type, temperature, fuel price, consumer price index (CPI), unemployment date, and Walmart promotional markdown event date. The challenge for this paper is to predict department-wide sales for each store and how does promotional markdown events (Super Bowl, Labor Day, Thanksgiving, and Christmas) affect the prediction, as Walmart decided that the holiday weeks are weighted five times higher than non-holiday weeks.

Dataset provided in Kaggle: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

There are five CSV files included in the dataset: Features, Stores, Train, Test. and SampleSubmission (will not be used, as it is for Kaggle submission only). Total of 16 attributes noted, and description of the attributes listed below (Table 1):

Attributes:	Description:
Store	This is the store number, it ranged from 1-45
Dept	It is the department number, it ranged from 1-99, for different categories of items
Date	The week date (all are Fridays)
IsHoliday	Whether that specific week has special holiday in it
Weekly_Sales	Store weekly total amount in USD
Temperature	Average weekly temperature of the specific store region in Fahrenheit
Fuel_Price	Cost of Fuel of the specific store region in USD
MarkDown1-5	Anonymized Data related to the promotional markdown of Walmart, the data is only available after November 2011, and not all store has it
CPI	Consumer Price Index of specific store region for the week
Unemployment	Unemployment rate of specific store region for the month
Type	Type of the store, A, B or C
Size	Size of the specific store measured in square feets

Attributes and Descriptions: Table 1

Below comprised the computation of descriptive statistic for numeric attributes (Table 2, Table 3):

	Weekly_Sales	Size	Temperature	Fuel_Price	CPI	Unemployment
min	(4,988.94)	34,875.00	(2.06)	2.47	126.06	3.88
max	693,099.36	219,622.00	100.14	4.47	227.23	14.31
range	698,088.30	184,747.00	102.20	2.00	101.17	10.43
median	7,612.03	140,167.00	62.09	3.45	182.32	7.87
mean	15,981.26	136,727.92	60.09	3.36	171.20	7.96
var	515,797,856.84	3,718,631,543.04	340.33	0.21	1,533.45	3.47
std.dev	22,711.18	60,980.58	18.45	0.46	39.16	1.86
SE.mean	34.98	93.92	0.03	-	0.06	-
CI.mean.0.95	68.56	184.08	0.06	-	0.12	0.01
coef.var	1.42	0.45	0.31	0.14	0.23	0.23
nbr.val	421,570	421,570	421,570	421,570	421,570	421,570
nbr.null	73	-	-	-	-	-
nbr.na	-	-	-	-	-	-

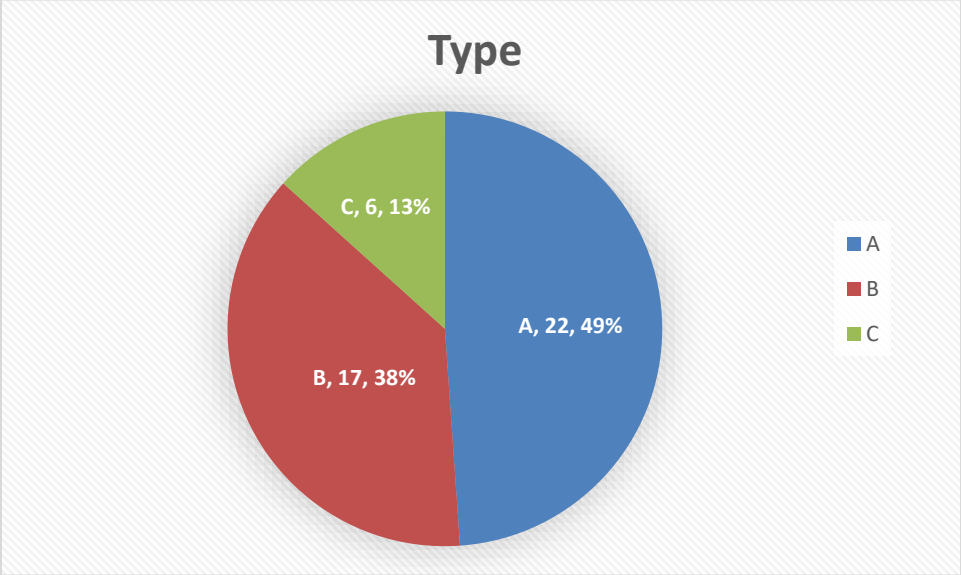
Descriptive Statistic for Number Attributes: Table 2

	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
min	0.27	(265.76)	(29.10)	0.22	135.16
max	88,646.76	104,519.54	141,630.61	67,474.85	108,519.28
range	88,646.49	104,785.30	141,659.71	67,474.63	108,384.12
median	5,347.45	192.00	24.60	1,481.31	3,359.45
mean	7,246.42	3,334.63	1,439.42	3,383.17	4,628.98
var	68,744,351.40	89,782,396.45	92,603,635.78	39,594,096.79	35,556,026.80
std.dev	8,291.22	9,475.36	9,623.08	6,292.38	5,962.89
SE.mean	21.36	28.41	25.99	17.13	15.32
CI.mean.0.95	41.86	55.68	50.94	33.57	30.03
coef.var	1.14	2.84	6.69	1.86	1.29
nbr.val	150,681	111,248	137,091	134,967	151,432
nbr.null	-	207	67	-	-
nbr.na	270,889	310,322	284,479	286,603	270,138

Descriptive Statistic for Number Attributes (Cont.): Table 3

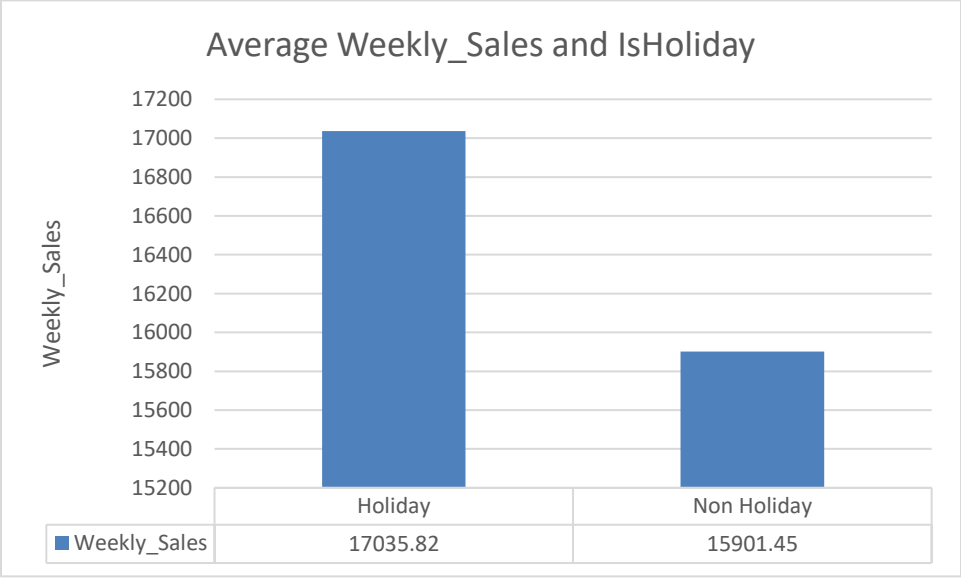
Some attributes that are Characters or Booleans, which are listed below:

Type: Based on the dataset, there are three types of stores, which is “A”, “B”, or “C”. In a total of 45 stores, almost half of the stores are marked as “A”, followed by “B” with 38% and “C” with 13% (Figure 1).



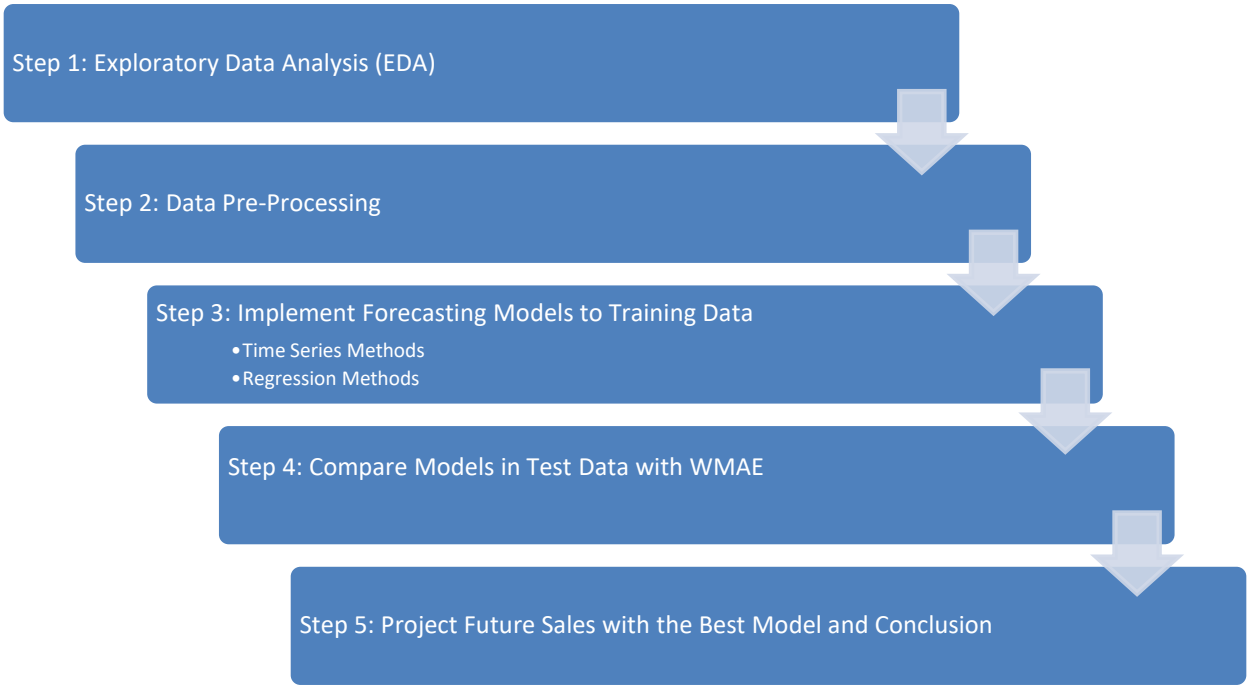
Type Attribute in Pie Chart: Figure 1

IsHoliday: We can also see that, based on the dataset, average weekly sales with holidays are higher.



Comparison of average weekly sales for holiday and non-holiday: Figure 2

Approach



Step 1: Exploratory Data Analysis (EDA)

After loading the dataset into R or Python, I would first explore the structure type and values of the dataset's CSV files. Then, generate a description statistic such as mean, median, minimum, maximum, and standard deviation for numeric attributes. Also, with the help of visual aid, a look at the frequency of category and Boolean attributes and explain the patterns or any findings regards to trends for holiday seasons and weekly sales. A Correlation matrix would also be performed to see any correlation between attributes.

Step 2: Data Pre-Processing

First, since different attributes are allocated in separated CSV files, I would be combining the datasets of Store.csv, Feature.csv, and Train.csv. There are N/As and null values in the dataset, therefore data cleaning, replacing null value, and impute missing value are needed. The column of Store and Dept would be merged to form a new attribute of Store_Dept (i.e. Store 1 and Department 1 become S1D1) for easier implementation.

Step 3: Implement Forecasting Models to Training Data

The data would be implemented into two separate approaches: Time Series and Regression. The training set and the testing set would be made with a ratio of 80/20.

For time series methods, I would first implement in traditional methods such as MA, Holt-Winters method, then ARIMA, SARIMA, VAR, then moving to deep learning methods such as LSTM and GRU.

For Regression approach, I would implement the model in univariate linear regression, multiple linear regression, then decision tree and random forest.

Step 4: Compare Models in Test Data with WMAE

Explanation of what is WMAE and how would it be computed in R. Also, a summary of the WMAE for all of the above methods would be used to compare against each other. A ranking table would be provided and show which method has the best score.

Step 5: Project Future Sales with the Best Model

The best model would be used to predict sales inside the test.csv file. Conclusion, limitation, and recommendation would also be stated for the different models and the dataset. Additional information such as which store and department were doing the best and which did poorly would be stated, so the management could target those specific departments and find what is the reason for high or low sales.

References

Baccar, Y. B., ROEUFF, F., LePennec, E., d'Alché-Buc, F., Bertrand, L. A. M. Y., & Jacques, D. O. A. N. (2019). Comparative Study on Time Series Forecasting.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019, December). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing* (pp. 462-474). Springer, Cham.

Bonnes, K. (2014). Predictive analytics for supply chains: A systematic literature review. In *21st twente student conference on IT. Netherlands*.

Goyal, A., Kumar, R., Kulkarni, S., Krishnamurthy, S., & Vartak, M. (2018). A solution to forecast demand using long short-term memory recurrent neural networks for time series forecasting. In *Midwest Decision Sciences Institute Conference*.

Hülsmann, M., Borscheid, D., Friedrich, C. M., & Reith, D. (2012). General sales forecast models for automobile markets and their analysis. *Trans. MLDM*, *5*(2), 65-86.

Jain, A., Menon, M. N., & Chandra, S. (2015). Sales Forecasting for Retail Chains.

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, *42*, 11-24.

Lasek, A., Cercone, N., & Saunders, J. (2016). Restaurant sales and customer demand forecasting: Literature survey and categorization of methods. In *Smart City 360°* (pp. 479-491). Springer, Cham.

Mentzer, J. T., & Moon, M. A. (2004). *Sales forecasting management: a demand management approach*. Sage Publications.

Pai, P. F., & Lin, C. S. (2005). Using support vector machines to forecast the production values of the machinery industry in Taiwan. *The International Journal of Advanced Manufacturing Technology*, *27*(1-2), 205.

Pai, P. F., Lin, K. P., Lin, C. S., & Chang, P. T. (2010). Time series forecasting by a seasonal support vector regression model. *Expert Systems with Applications*, *37*(6), 4261-4265.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 15.

Udom, P., & Phumchusri, N. (2014). A comparison study between time series model and ARIMA model for sales forecasting of distributor in plastic industry. *IOSR Journal of Engineering*, *4*(2), 32-38.

Wu, C. S. M., Patil, P., & Gunaseelan, S. (2018, November). Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 16-20). IEEE.